DCCUMENT RESUME

ED 050 148 TM 000 536

AUTHCR Ebel, Robert L.
TITLE The Comparative Effectiveness of True-False and Multiple Choice Achievement Test Items.
PUB DATE Feb 71
NOTE 5p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971
AVAILABLE FROM Robert L. Ebel, 449 Erickson Hall, Michigan State University, East Lansing, Michigan 48823

EDRS PRICE EDRS Price MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS *Achievement Tests, Comparative Analysis, *Item Analysis, *Multiple Choice Tests, *Test Construction, *Test Reliability, Test Selection
IDENTIFIERS *True False Tests

ABSTRACT

The suggestion that multiple-choice items can be converted to true-false items without essentially changing what the item measures and with possible improvement in efficiency is investigated. Each of the 90 four-choice items in a natural science test was rewritten into a pair of true-false items, one true, one false. The resulting 180 items were divided to make two 90-item forms A and B which were administered to chance halves of a class of college students. Following item analysis, the most highly discriminating member of each pair of items was chosen for further comparison with the multiple choice forms. Using these selected T-F items, two additional experimental forms, half true-false and half original multiple choice items, were then constructed and administered. Analysis of the resulting data indicates that true-false test items, item for item, are less discriminating than multiple-choice items. This gives partial support to the belief that minute for minute a true-false test can be as reliable as a multiple choice test. It also indicates some support to the hypothesis that there is no important difference in what the two item forms measure. Overall results, despite their limitations, tend to strengthen rather than weaken faith in the usefulness and value of true-false test items. (LR)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

The Comparative Effectiveness of True-False and Multiple Choice Achievement Test Items

Robert L. Ebel

Michigan State University

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY

Robert Ebel

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

ED050148

TM 000 536

1. Reason for the study

Although many test specialists hold true-false test items in low esteem, a few see special virtues of efficiency and ease of preparation in them and advocate their wider use.[1] One of the arguments advanced in their behalf is that multiple choice test items can be converted to true-false items without changing what the item measures in any important way, and with possible improvement in efficiency (reliability per hour of testing time). This study was designed to yield data that might support or weaken that argument.

2. Procedures of the study

An expertly constructed, highly regarded published test of natural science was chosen as the starting point of the inquiry. The test consists of 90 four-alternative multiple choice items. It is intended for use by the general population of high school students in any of the high school grades.

The investigator converted each of the multiple choice items into a pair of true false items, one true, the other false, both intended to test essentially the same understanding as the original multiple choice item. Exhibit 1 shows an item in the original multiple choice form and in revised true-false form.

The resulting 180 true false items were divided into two 90 item forms, A and B, so that one member of each pair would appear in each of the two forms. Forms A and B were administered to chance halves of a class of 65 students enrolled in an introductory college level course in testing and grading. While these were not the kind of students for which the test was originally written, their understanding of natural science was not so much better, or so much more uniform, than that of typical high school students as to impair the usefulness of their responses for item analysis.

The most highly discriminating member of each pair was then chosen for further study in comparison with the multiple choice forms. Using the selected true-false items and the original multiple choice items, two additional experimental forms of the test of natural science were constructed. In each of the first 44 items were multiple choice items. The next 44 were

---

1. Robert L. Ebel, Measuring Educational Achievement. Englewood Cliffs, New Jersey Prentice-Hall, Inc. (1965) Chapter 5.

true-false items over different concepts. The limitation of the half tests to 44 items, rather than 45 was a concession to convenience in obtaining sub-test scores from a four-column answer sheet. Form C consisted of 44 of the 45 odd-numbered multiple choice items from the original test, plus 44 true-false items derived from the even-numbered multiple choice items of the original test. Form D was the complement of Form C, using even-numbered multiple choice items first, and then true-false items based on the original odd numbered items. Forms C and D were administered to chance halves of a class of 102 students enrolled in an introductory college level course in testing and grading.

3. Results of the data analysis

Table 1 presents the item composition of the two tryout and two final forms, statistics of the score distributions, and measures of the item discrimination and test discrimination (reliability). None of these data bear directly on the question being investigated. They are presented for background information.

Table 2, however, presents data bearing directly on the point at issue. It compares the multiple-choice and true-false sections of the two final forms. In Form C the mean index of discrimination (Mean D) of the true-false items (.30) is only a little less than that for the multiple choice items (.33). In Form D the true-false items looked much worse (Mean D = .17) than the multiple choice items (Mean D = .38). The differences in item discrimination are reflected in corresponding differences in score reliability (K.R. 20).

Table 1. Data on Test Forms

	Tryout		Final	
Forms	A	B	C	D
Items				
Multiple-Choice	0	0	1-44	1-44
True-false	1-90	1-90	45-88	45-88
Scores				
n	32	34	53	50
Mean	67	63	62	63
S.D.	7.7	8.1	11.1	9.7
Discrimination				
Mean D	.21	.21	.31	.28
K R 20	.77	.77	.88	.85

Table 2. Data on Item Forms

Test Form	C	C	D	D
Item Type	MC	TF	MC	TF
Item Numbers	1-44	45-88	1-44	45-88
Scores				
n	53	53	50	50
Mean	33.1	28.8	32.3	30.7
S.D.	5.86	5.42	6.85	4.08
Discrimination				
Mean D	.33	.30	.38	.17
K R 20	.81	.72	.86	.55
Adjusted K.R. 20		.84		.71
Correlation				
(MC-TF)	.92		.55	
Corrected for atten.	1.20		.80	

The adjusted K.R. 20 values were obtained by applying the Spearman Brown formula to predict the reliability of an 88 item true false test. The rationale for such an adjustment is that students typically answer two true-false items, or more, in the time required to answer one multiple choice item such as those used in this study. In Form C the adjusted true-false reliability (.84) is slightly higher than that of the multiple choice items (.81). However in Form D even the adjustment fails to bring the true-false reliability (.71), close to that of the multiple choice items (.86).

The bottom section of Table 2 presents the correlations between scores on multiple-choice and true-false items in Form C (.92) and Form D (.55). When corrected for attenuation these correlations become 1.20 and .80 respectively. Note that the mean of the corrected values is 1.00.

4. Interpretation of the results

These data confirm the expectation that item for item true false test items tend to be less discriminating than multiple choice items, though in some cases the difference is surprisingly small. They give partial support to the belief that minute for minute a true-false test can be as reliable as a multiple choice test. They also give some support to the hypothesis that there is no important difference in what the two item forms measure. Overall the correlation between sub test composed of the two forms is as high as their reliabilities will allow.

We have no good explanation other than sampling fluctuations, for the differences observed between Form C and Form D. Clearly it would have been much better to have had n's of 300 for each of the final forms. It also would have been better if the selection of items from the tryout forms could have been based on more responses than those provided by 32 or 34 examinees. Table 3 shows how much the indices of discrimination for the same item varied from tryout to final forms. Values for the same item are circled. Table 3 also shows the low correlation between indices of discrimination for the "same" item in true-false and multiple-choice form. Most of these differences are probably attributable to instability (sampling errors) in the indices themselves.

When the study is repeated we should, in addition to using much larger n's, use the true-false tryout data less mechanically. With more stable indices as a basis from which to work, we should do more revision of the true-false items, and seek qualitative as well as quantitative bases for the final selection. The multiple choice items against which the true-false items were being compared were given a much more adequate tryout and much more extensive and careful revision.

Table 3. Discrimination Indices for Related Items Based on the Same Content

Item Number	Tryout Forms*		Final Forms**	
	A	B	T-F	M-C
1	37	(56)	(38)	07
2	12	(34)	(35)	31
3	(62)	00	(23)	14
4	(37)	22	(-07)	31
5	13	33	(31)	29
6	(12)	-11	(21)	15
7	12	(22)	(07)	14
8	(25)	-11	(22)	38
9	(50)	22	(31)	07
10	-25	(89)	(22)	23

* 90 True-false items in each form, A and B

** 44 Multiple-choice items and 44 true false items in each form, C and D. Form C included the odd-numbered multiple-choice items and the even-numbered true-false items. Form D included the others.

Overall, however, despite their serious limitations, the results of this study do more to strengthen than to weaken faith in the usefulness and value of true-false test items.

Exhibit 1. Sample Items

A. Original multiple choice form

1. What enables man to live in a greater range of climates than most other animals
   1. He is stronger than other animals
   2. He is a warm-blooded animal
   *3. He can control his surroundings to a greater extent
   4. He eats less than other animals

B. Alternative true-false forms

A. Man can live in a greater range of climates than most other animals because he is warm-blooded. F

B. Man is less dependent on his immediate environment for food and comfort than are most other animals. T